1.选题意义与应用背景

随着互联网的普及和Wifi技术的发展，无线/移动通信网络已经成为当今社会的基础设施，广泛地应用于政府、金融、军事等领域。统计显示，截止到2017年12月，中国网民规模达7.72亿，互联网普及率达到55.8%【2018年第41次中国互联网络发展状况统计】；与此同时，随着支付宝和微信支付等线下手机支付方式的普及，我国网民在线下消费中使用手机支付比例由2016年的50.3%提升至65.5%。无线网络设备的普及在给人们生活带来便利的同时，也存在很多安全隐患，如隐私泄露、病毒入侵等，因此对网络安全的防御和无线网络信息安全的需求日益迫切。

传统的保护无线网络安全的方法通常是基于比特层面的(即OSI七层模型中物理层以上的层次),通过设计基于密码机制的安全协议来实现对数据完整性和机密性的保护以及提供通信双方身份的认证【无线通信设备的射频指纹提取与识别方法】。然而,密码机制存在密钥泄露的风险，无线网络安全协议也通常会存在安全漏洞【基于射频指纹的无线网络物理层认证关键技术研究】。例如,IEEE 820.11 无线局域网(WLAN)最初的有线等效加密(WEP)协议易受统计分析攻击【同上，2】,虽然此后升级为WPA和WPA2,但其口令句可以被恢复,仍然存在着各种各样的安全问题【同上，3】。此外,一旦密钥泄露，现有的安全机制无法实现身份认证。因此,人们急需寻找一种新型的安全机制有效识别授权用户和非授权用户,从而降低来自恶意用户的潜在威胁。值得庆幸的是，攻击者可以通过多种方式获取用户的账号密码，甚至模拟用户的操作，但却很难伪造出与用户相似的设备特征。基于此，研究人员提出可以在密钥认证的基础上，根据用户设备的特征再次认证，从而准确识别用户。这样即使用户的账号被盗，攻击者仍然无法登陆用户的账号，极大地增强了用户的信息安全。

这种技术的重要应用有：入侵检测、身份认证、司法数据收集【Device Identification via Analog Signal Fingerprinting: A Matched Filter

Approach】

2.相关工作

“指纹”最早是指标识个体的生物特征识别技术【**Accelprint\_NDSS14**，Information fusion in biometrics】【P. Tuyls and J. Goseling,“Capacity and examples of templateprotectingbiometric authentication systems,”】。这一概念早在上世纪60年代就被应用于设备识别，当时研究者们开发出可观测到信号特征的系统来区分雷达【K. Talbot, P. Duley, and M. Hyatt, “Specic emitter identification andverification,”】。此后类似的技术被用于蜂窝网络中的发射器识别【M. Reizenman, “Cellular security: better, but foes still lurk,”】【L. Langley, “Specific emitter identification (sei) and classical parameter fusion technology,”】。近年来，随着对无线网络设备研究的深入，对无线/移动网络设备“指纹”的分析技术也越来越受到研究者的重视，展现了广阔的应用前景【】。

设备指纹是指可用于唯一标识出该设备的设备特征或者独特的设备标识。目前较普遍的技术是针对设备提供的某些信息，从中提取特征生成近似唯一的设备指纹，与存储的可信信息匹配辨识，从而对设备的身份进行认证和辨识。其原理是不同厂商生产的设备（即便是同一厂商生产的设备）会由于硬件或软件的不一致在某些特定的方面存在些许差异。当前研究的技术手段可大致分为基于硬件的识别和基于软件的指纹识别。基于软件的识别通常是从流量信息中获取特征指纹，而基于硬件的识别则往往是根据设备的传感器信息或时钟偏移等硬件设备的信息对设备进行标识。

2.1基于软件的设备识别

802.11 MAC帧的格式和内容经常被用于辨识无线设备。MAC地址是Medium/Media Access Control地址的简称，表示互联网上每一个站点的标识符。在网络底层的物理传输过程中，是通过物理地址来识别主机的，所以MAC地址就如同居民身份证上的身份证号码一样，具有全球唯一性。Guo等人根据MAC帧头的信息追踪设备【Do You Hear What I Hear? Fingerprinting Smart Devices Through Embedded Acoustic Components，参考文献40】。顾杨等人利用从MAC层中的管理帧提取的特征指纹作为区分真假AP的特征指纹【】。他们将特征指纹又分为基于无线端的特征指纹和基于特种MAC帧刺激响应的特征指纹，其中基于无线端的特征指纹主要从信标帧和探测响应帧中提取，而基于特种MAC帧刺激相应的特征指纹主要从探测响应帧、关联响应帧和认证帧中提取。第一种特征指纹是通过被动的扫描无线网络获得的，第二种特征指纹是利用构造特殊格式MAC帧主动刺激AP（或无线路由器），捕获AP（或无线路由器）响应获得的。实验结果表明这两类无线设备的特征指纹可以检测出目前大部分无线钓鱼AP。但是，MAC地址是与网卡对应的，如果一个设备拥有不止一个网卡，这个设备就会有相应数目的MAC地址，此时MAC就不能用来唯一地标识设备；另外，如同居民的身份证容易被人盗用或者伪装，设备的MAC地址也容易被篡改或者伪装。

鉴于使用MAC的识别技术中存在上述缺陷，研究者们提出了一系列根据其他特征构建指纹的方法。Desmond等人仅通过研究802.11的请求探测帧，即可识别出连接在同一无线接入点（WLAN:）上的设备。【Do You Hear What I Hear? Fingerprinting Smart Devices Through Embedded Acoustic Components ，参考文献30】。Loh等人通过分析无线局域网（WLAN：Wireless Local Area Networks）中不同设备对802.11请求探测帧响应的时序特性，从中提取相应的指纹【Identifying Unique Devices through Wireless Fingerprinting】。Pang等人也能从流量中提取出特征指纹，进行设备的认证与标识【Do You Hear What I Hear，参考文献57】。Seika【Active Fingerprinting of 802.11 Devices by Timing Analysis】等人向802.11设备发送一个信号，捕获响应帧的到达时间并用时序分析研究其规律，用SVM分类器建立目标设备模型。~~他们在5台设备上测试该方法并达到86%的平均精度~~。

Gao K等人【A passive approach to wireless device fingerprinting】提出了一种用于确定接入到网络的AP的类型的基于黑盒的被动式识别技术。作者进行了大量的实验（收集超过60GB的数据）以对6种AP进行分类，在至少100000个数据包的数据基础上取得了较高的分类精度。无线网络中对AP的识别通常利用网络协议中常见的标识符，例如网络名称、MAC地址或者IP地址，但是这些标识符很容易被伪造、拦截或者更改。

此外，一些开源的工具如Nmap【Do You Hear What I Hear，50】和Xprobe【Do You Hear What I Hear，68】可以通过TCP/IP协议栈的响应识别设备的操作系统。其他基于软件的指纹识别技术则是根据设备上的应用如浏览器来确定设备。Yen等人【同上69】曾经分析长达一个月的必应和Hotmail的日志，成功追踪到用户。Acar等人通过研究JavaScript和一些较流行的第三方插件如Flash Player来获取设备上的字体列表，以此追踪用户【同上，18】。其他研究人员提出使用性能基准来区分JavaScript引擎【Fingerprinting Information in JavaScript Implementations】，浏览器的浏览历史也被用于追踪web端的用户【同上，56】。

2.2基于硬件的设备识别

Bratus等人【Active behavioral fingerprinting of wireless devices】提出了一种主动式设备指纹识别方法,该方法通过向802.11无线设备发送一系列的经过特殊构造的某种错误格式的数据帧来观察设备的回应，通过这些回应的差异可以区分设备在芯片、固件或者驱动上的差异，以此达到设备识别的目的。同时，该方法也可以借助无线网卡芯片或无线网卡驱动的特征指纹信息来区分无线钓鱼AP和合法授权AP，为以较低成本来部署主动式无线端的检测方法提供了一些思路。

Radhakrishnan等人设计了一种名为GTID的识别框架，该技术采用主动和被动结合的方式，不仅可识别物理设备，还可以判断出相应的设备类型【GTID：A Technique for Physical Device and Device Type Fingerprinting】。GTID实现的主要依据是设备之间的异构性，他们认为不同的设备的处理器、DMA控制器和时钟偏移等都存在差异，而这些差异可以被用于识别设备和设备类型。GTID通过被动地抓取网络流量，从中提取出相应的指纹特征并使用人工神经网络（ANN）的算法进行训练和测试，取得了良好的效果。他们的技术适用于各种网络协议，且不需要进行深度包检测，但是由于该技术依赖于细粒度包时间，在路由器和交换机的缓冲区有时间丢失，故该技术的还有待进一步的研究。

基于硬件的指纹识别技术依赖于某些稳定的特性。【Estimation and removal of clock skew from network delay measurements】中的研究表明网络设备往往具有稳定的时钟偏移，Kohno和Cristea等人的研究便是基于此，他们通过分析TCP和ICMP中微小的时钟偏移，从而构建设备的特征指纹【Remote Physical Device Fingerprinting.被动】【Fingerprinting Smartphones Remotely via ICMP Timestamps】。该技术无需对设备做任何修改，即便目标设备通过多种方式接入互联网，或是测量设备与目标设备相距数千英里，均能达到较好的识别结果。

然而，时钟偏移率很大程度上依赖于实验环境【Specific emitter identification (SEI) and classical parameter fusion technology】。【Clock Skew Based Remote Device Fingerprinting Demystified】中作者研究了基于时钟偏差的无线设备识别的局限性，他们利用无线接入点（AP：Access Point）在信标帧中定期发送的时间戳为依据进行识别，消除了测量设备对时钟偏移的影响。此外，他们还进行了大量的评估，以探讨不同接入点和测量装置之间时钟偏差的分布及稳定性，发现时钟偏差的波动仅为1ppm。该算法能够消除测量装置对实验结果的影响，使不同设备生成的指纹可比较且具有区分性。

与仅使用时钟偏斜作为特征指纹不同的是，Neumann C等人【An empirical study of passive 802.11 Device Fingerprinting】分别以传输速率、帧大小、介质访问时间、传输时间、包内间隔时间等特征作为特征指纹进行802.11设备指纹识别，并比较这些特征在设备识别中的性能差异，为其他的设备指纹识别工作提供了很好的借鉴。Franklin等人研究了在IEEE 802.11兼容设备上运行的固件和设备驱动程序的差异，在设备未被更改的情况下可有效精确地识别无线驱动程序【Passive Data Link Layer 802.11 Wireless Device Driver Fingerprinting】。

Gerdes等人提出了基于模拟信号的终端设备识别技术，最多仅需要分析25帧就能高效识别设备。这种指纹识别技术是基于设备的制造和硬件组件在数字领域创造的独特的信号特征。【Device Identification via Analog Signal Fingerprinting】。虽然实验结果较好，但他们的研究仅是针对有线设备而言，这种方法对无线网络设备是否可行还存在疑问。而且这种方法依赖于如模数转换器和数字取样示波器等昂贵的设备，因此实用性不高。

此外，还有许多人研究了利用设备的辐射测量量作为指纹的技术【Wireless device identification with radiometric signatures】【Device Fingerprinting to Enhance Wireless Security using Nonparametric Bayesian Method】【Securing Wireless Systems via Lower Layer Enforcements】【Improved radiometric identification of wireless devices using mimo transmission】。其原理是天线、功率放大器、ADC、DAC等硬件在生产过程中不可能完全相同，每个设备因此会具有一系列独一无二的辐射信号，如振幅、频率、相位等，从中可提取出特征指纹。由于这些辐射测量量在设备生产后无法改变，所以使用起来安全可靠。这种方法最大的局限性在于数据采集困难，只能在有限的范围内获取到设备的指纹，无法进行远程追踪或认证。

Anupam Das【Do You Hear What I Hear? Fingerprinting Smart Devices Through Embedded Acoustic Components】等人认为厂商在制造智能手机的麦克风和扬声器时存在不同程度的缺陷，通过分析不同智能手机上的麦克风和扬声器的声学特征可以从中提取到设备相应的指纹。他们的方法虽然在50个安卓手机上达到了98%的精度，但是必须在手机的麦克风或扬声器发声的时候才能进行测试，实现复杂且无法远程监控。

在论文【AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable】中，作者认为智能手机和平板电脑中的加速度计具有独特的指纹，可以用于设备识别。他们在25个安卓手机和两个平板电脑上验证这一方法，识别精度达到96%。这种方法虽然精度高，但是要求设备上必要要有加速度传感器，且需要某种形式的外部刺激/震动来捕获加速度计的数据，局限性较大。相比而言，本文的工作只需远程捕获网络流量数据，不会被用户发现，操作简单，可实现性高。

3.流量数据帧介绍

本文中我们使用的数据流量帧为通过Wireshark抓包软件或jnetPcap API获取的数据流量帧。以图 2‑2为例介绍数据帧的格式。

* **物理层：**

Frame 239823 (61 bytes on wire, 61 bytes captured) 具体地，代表数据帧的帧号和实际捕获的数据字节数；

Arrival Time: Jan 21, 2008 15:17:33.910261000 捕获数据帧的具体日期和时间；

[Time delta from previous packet:0.00000 seconds] 本数据帧与前一帧的间隔时间；

[Time since reference or first frame: 0.00 seconds] 本数据帧与第1帧的间隔时间  
Frame Number: 239823 帧序号；  
Packet Length: 61 bytes 帧长度；  
Capture Length: 61 bytes 捕获长度；  
[Frame is marked: False] 此数据帧是否做了标记：否；  
[Protocols in frame: eth:ip:tcp] 数据帧内所封装的协议层次结构；  
[Coloring Rule Name: HTTP] 用不同颜色染色来标记的协议名称：HTTP  
[Coloring Rule String: http || tcp.port == 80] 染色显示规则的字符串：

* **数据链路层：**

Ethernet II, Src: AcerTech\_5b:d4:61 (00:00:e2:5b:d4:61), Dst: Jetcell\_e5:1d:0a (00:d0:2b:e5:1d:0a) 以太网协议版本：II，源MAC地址，目的MAC地址；

Destination: Jetcell\_e5:1d:0a (00:d0:2b:e5:1d:0a) 目的MAC地址：厂名\_序号（网卡地址）；  
Source: AcerTech\_5b:d4:61 (00:00:e2:5b:d4:61) 源MAC地址：厂名\_序号（网卡地址）；  
Type: IP (0x0800) 数据帧内封装的上层协议类型为IP（十六进制码0800）；

* **网络层：**

Internet Protocol, Src: 202.203.44.225 (202.203.44.225), Dst: 202.203.208.32 (202.203.208.32) 网络层协议，源IP地址，目的IP地址；

Version: 4 网络层协议IPv4；

Header length: 20 bytes IP包头部长度；  
Differentiated Services Field:0x00(DSCP 0x00:Default;ECN:0x00)差分服务字段；  
Total Length: 48 IP包的总长度；

Identification:0x8360 (33632) 标志字段

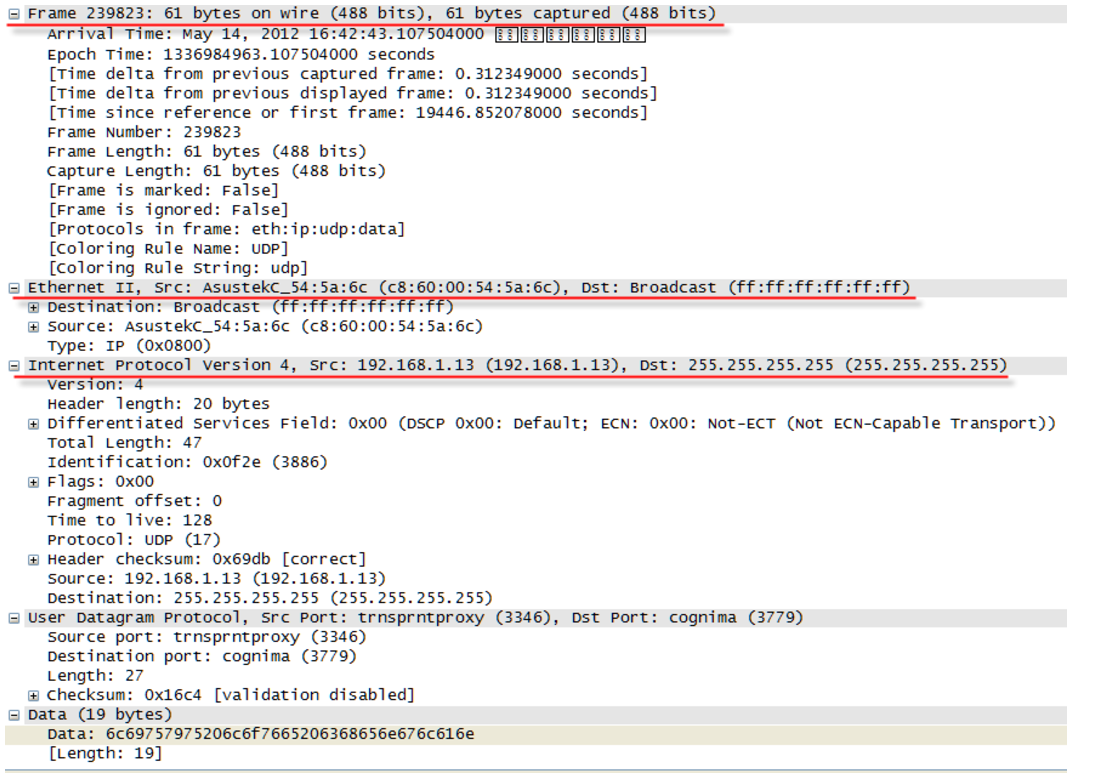
Flags: 标记字段，在路由传输时，判断是否允许将此IP包分段；

Fragment offset: 0 分段偏移量，将一个IP包分段后传输时，再次重组的标识；  
Time to live: 128 存活期TTL；  
Protocol: TCP (0x06) 此包内封装的上层协议为TCP;  
Header checksum: 0xe4ce [correct] 头部数据的校验和;  
Source: 202.203.44.225 (202.203.44.225) 源IP地址;  
Destination: 202.203.208.32 (202.203.208.32) 目的IP地址;

* **传输层：**

Transmission Control Protocol, Src Port: 2764 (2764), Dst Port: http (80), Seq: 0, Len: 0 传输控制协议TCP的内容;  
Source port: 2764 (2764）源端口名称（端口号）;  
Destination port: http (80) 目的端口名http（端口号80）;  
Sequence number: 0 (relative sequence number) 序列号（相对序列号）;  
Header length: 28 bytes 头部长度;  
Flags: 0x02 (SYN) TCP标记字段（本字段是SYN，是请求建立TCP连接）;  
Window size: 65535 流量控制的窗口大小;  
Checksum: 0xf73b [correct] TCP数据段的校验和;  
Options: (8 bytes) 可选项;

由上述数据帧的结构可以看出，数据帧中包含了设备在访问互联网时的各种时空信息，为本文基于数据帧进行设备行为特征身份识别奠定了基础。



4.分类算法介绍

4.1随机森林

随机森林（RF：Random Forests）是一种比较新的机器学习模型。上世纪八十年代Breiman等人提出分类与回归树的算法（CART：Classification and regression tree）【Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees[M]// Classification and regression trees. Wadsworth International Group, 1984:17–23.】，通过反复二分数据进行分类或回归，将输入空间即特征空间划分为有限个单元，并在这些单元上确定预测的概率分布。

2001年Breiman又将分类树组合成随机森林【Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.】。随机森林由很多的决策树组成的，这些决策树之间互相没有关联。形成森林之后，当有一个新的输入样本进入的时候，每个决策树测试待分类项中相应的特征属性（这些属性可以是分类项不同的属性），并按照判定条件将输入样本归类到决策树不同的分支，直到最后到达叶节点，叶节点即为决策树预测的类。最后看哪个类被选择的最多，就预测这个样本属于哪个类。如图2‑4，要对一个人的身份进行二分类判定，不同的决策树就可以分别根据该人的年龄、是否是学生、信用评级等身份属性来判定，最后根据投票结果进行投票表决，示例中，五个叶节点，三个判定为“是”，两个判定为“不是”，所以对这个人的身份作肯定的判定。

随机森林的“随机”体现在以下两个方面：1）在训练每棵树时，从训练样本池（N个样本）中选取的训练子集（n个样本，n<N）是随机有放回地选取的。这样在训练时，每一棵树的输入样本都不是全部的样本，可以避免过拟合现象；2）在决策树的每个节点进行分裂时，从输入特征（M个特征）中随机选取m个输入特征（m<M）,然后从这m个特征里选一个最好的进行分裂。与分类回归树算法相比，随机森林在运算量没有显著增加的前提下提高了预测精度。



图 2‑4随机森林示例

4.2支持向量机

支持向量机（SVM：Support Vector Machine）是Cortes和Vapnik在1995年提出的经典分类算法,它在解决小样本、非线性及高维模式识别中表现出许多特有的优势【小贾毕设，15】。SVM是一种有监督式学习的分类模型，它分类的思想是：给定一个包含正例和反例的训练样本集合，SVM找到一个超平面对样本根据正例和反例进行分割（如图2-5），超平面应当位于正例和反例之间且距离两类数据样本的距离尽可能远。SVM算法的核心在于找到一个核函数，使得在低维空间中线性不可分的问题在高维空间中变得线性可分，这样做的好处是降低在原数据空间进行非线性曲面分割计算的复杂度。

选择一个恰当的核函数不仅可以在很大程度上降低SVM分类的时间复杂度，还可以提高SVM的分类精度，如何根据实际的训练数据选择恰当的核函数是SVM在应用中亟待解决的一个关键问题。目前常用的SVM核函数有：线性核函数、多项式核函数、Sigmoid核函数等。

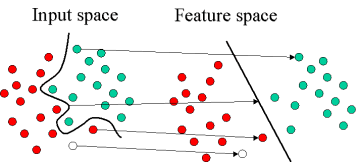


图 2‑5 支持向量机示例

4.3朴素贝叶斯

贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类。而朴素贝叶斯分类(NBC：Naïve Bayes Classifier)是贝叶斯分类中最简单却应用最为广泛的一种方法。朴素贝叶斯分类器基于一个简单的假设：给定样本的特征之间相互条件独立。根据贝叶斯定理，给定样本特征x,该样本属于类别y的概率如式(2-1)所示：

 (2-1)

式中，x是一个特征向量，将其维度设为M。根据朴素贝叶斯样本特征之间相互独立的假设，式(2-1)可以用式(2-2)表示：

 (2-2)

通过公式2-2，求样本特征x属于类别y的概率就转化为统计类别y的先验概率问题。在给定已知的训练样本集合以及类别标签的条件下，计算出未知的样本属于每一类别的条件概率，取概率最大的一类为最终分类结果。

朴素贝叶斯分类算法逻辑简单，易于实现，且分类过程中时空开销小。理论上，朴素贝叶斯模型与其他分类算法相比有最小的误差率。但事实却并非总是如此，这是因为朴素贝叶斯模型假设特征之间相互条件独立，这个假设在实际应用中很多时候并不成立。因此在特征数目较多或特征之间相关性较大时，分类效果不好；而在属性相关性较小时，朴素贝叶斯分类的性能最为良好。

#### K最近邻法

K最近邻（KNN）分类算法是最简单的监督式算法之一。如图 2‑6所示，它的分类思想是从训练数据集中找到和新数据最接近的k条数据，这k条数据中的大多数样本属于某个类别，待分类样本就属于这个类别。KNN的计算步骤为1）计算距离：给定待分类数据，计算该数据与训练集中每条数据的距离；2）寻找邻居：寻找与待分类数据距离最近的k个训练数据，作为该数据的最近邻；3）进行分类：确定这k个最近邻所属的主要类别，来判定待分类数据所属的类别。KNN算法简单，易于理解，易于实现，适合对稀有事件分类和处理多分类问题，但是参数k的取值是一个问题，k太小时，分类结果易受噪声点影响；k太大时，最近邻点中又可能包含太多其他类别的点。本文的实验中，我们将k取值为30。

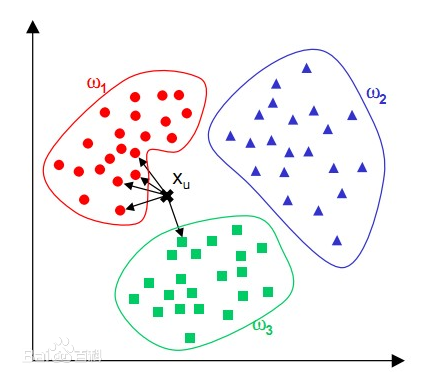


图 2‑6 k最近邻示例

#### 人工神经网络

ANN是启发于生物学上的神经网络的计算模型，从结构和功能上都与生物神经网络类似。一个ANN由一组名为神经元的互联计算模型所组成，这些神经元根据特定的激活函数将输入转换为输出。虽然在我们的工作中仅将ANN视为一个黑盒，但是我们主要使用由一个输入层、一个隐含层和一个输出层组成的前馈神经网络。该类ANN通常需要监督式学习用于预测、模式识别和非线性函数拟合。在我们的ANN中，使用共轭梯度反向传播作为训练函数，ANN输出0-1之间的值，其中1代表完美匹配。

图 2‑7显示了一个ANN的示意图，这个ANN可使用含有N个窗口的指纹从而将M个不同的设备或设备类型进行分类，这是一个包含一个输入层、一个隐含层和一个输出层的多层前馈ANN。输入层接受大小为N的向量（）,并产生一个大小为M的输出向量（）。输入向量的元素对应于概率密度（签名）中的值，并且输出向量的元素对应于输入签名与被训练的M个设备或设备类型签名之间的相似性度量。

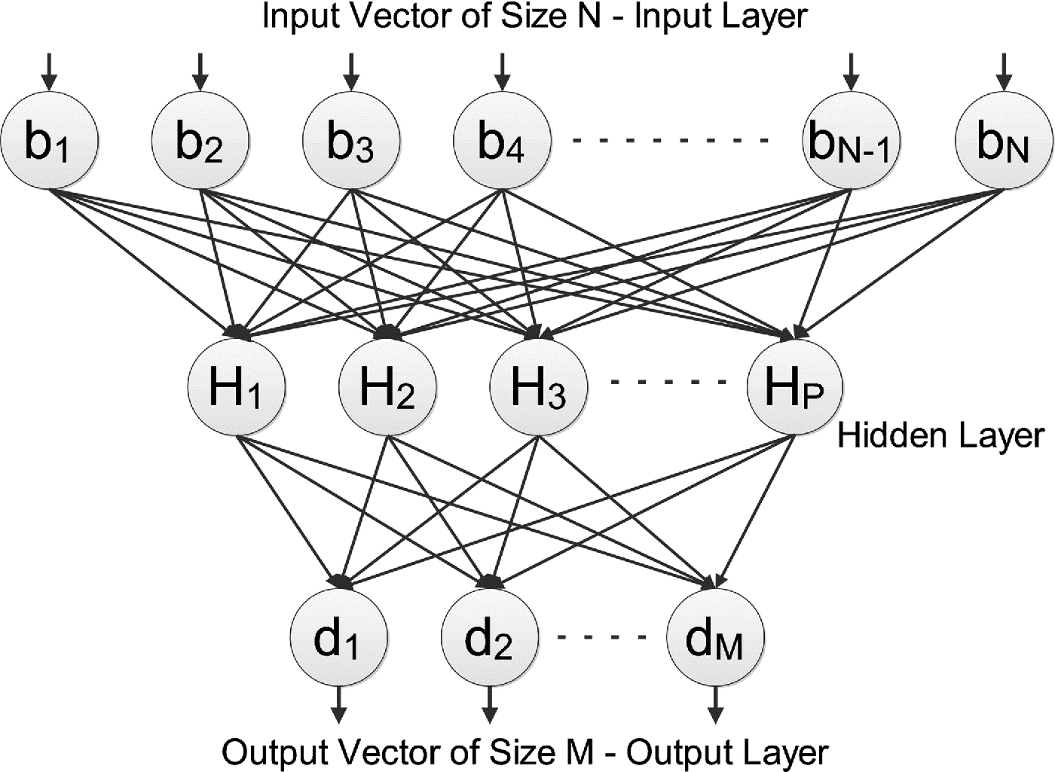


图 2‑7简单的神经网络示例